METHOD FOR DEVELOPING A CLASSIFIER FOR CLASSIFYING COMMUNICATIONS

BACKGROUND

[0001] The present invention is a computer assisted/implemented tool that allows a non-machine learning expert to build text classifiers. The present invention is also directed to the task of building Internet message relevancy filters.

[0002] The full end-to-end process of building a new text classifier is traditionally an expensive and time-consuming undertaking. One prior approach was to divide the end-to-end process into a series of steps managed by people with different levels of expertise. Typically, the process goes as follows: (1) a domain expert/programmer/machine-learning expert (DEPMLE) collects unlabeled communications (such as, for example, text messages posted on an Internet message board); (2) the DEPMLE writes a document describing the labeling criteria; (3) hourly workers with minimal computer expertise label a set of communications; (4) a data quality manager reviews the labeling to ensure consistency; and (5) the DEPMLE takes the labeled communications and custom-builds a text classifier and gives reasonable bounds on its accuracy and performance. This process typically takes several weeks to perform.

[0003] Traditional text mining software simplifies the process by removing the need for a machine learning expert. The software allows a tool expert to provide labeled training communications to a black box that produces a text classifier with known bounds on its accuracy and performance. This approach does not cover the complete end-to-end process because it skips entirely over the cumbersome step of collecting the communications and labeling them in a consistent fashion.

[0004] The traditional approach for labeling data for training a text classifier presents to the user for labeling, sets of randomly-selected training communications (un-labeled communications). Some of the user-labeled communications (the "training set") are then used to "train" the text classifier through machine learning processes. The rest of the user-labeled

communications (the "test set") are then automatically labeled by the text classifier and compared to the user-provided labels to determine known bounds on the classifier's accuracy and performance. This approach suffers in two ways. First, it is inefficient, because better results can be achieved by labeling smaller but cleverly-selected training and test sets. For example, if a classifier is already very sure of the label of a specific unlabeled training example, it is often a waste of time to have a human label it. The traditional approach to solving this problem is called Active Learning, where an algorithm selects which examples get labeled by a person. The second problem with human labeling is that it is inaccurate. Even the most careful labelers make an astonishingly high number of errors. These errors are usually quite pathological to training a classifier. For example, when building message relevancy filters, a very significant fraction of time may be spent relabeling the messages given by a prior art Active Learning tool.

SUMMARY

[0005] The present invention is directed to a computer assisted/implemented method for developing a classifier for classifying communications (such as text messages, documents and other types of communications, electronic or otherwise). While the exemplary embodiments described herein are oriented specifically toward the task of building message relevancy filters, the present invention also provides a framework for building many types of classifiers. The present invention is further directed to a computer or computer system (or any similar device or collection of devices) operating a software program including instructions for implementing such a method, or to a computer memory (resident within a computer or portable) containing a software program including instructions for implementing such a method.

[0006] Use of the computerized tool according to the exemplary embodiment of the present invention comprises roughly four stages, where these stages are designed to be iterative: (1) a stage defining where and how to harvest messages (i.e., from Internet message boards and the like), which also defines an expected domain of application for the classifier; (2) a guided question/answering stage for the computerized tool to elicit the user's criteria for determining whether a message is relevant or irrelevant; (3) a labeling stage where the user examines

carefully-selected messages and provides feedback about whether or not it is relevant and sometimes also what elements of the criteria were used to make the decision; and (4) a performance evaluation stage where parameters of the classifier training are optimized, the best classifier is produced, and known performance bounds are calculated. In the guided question/answering stage, the criteria are parameterized in such a way that (a) they can be operationalized into the text classifier through key words and phrases, and (b) a human-readable English criteria can be produced, which can be reviewed and edited. The labeling phase is heavily oriented toward an extended Active Learning framework. That is, the exemplary embodiment decides which example messages to show the user based upon what category of messages the system thinks would be most useful to the Active Learning process.

[**0007**] The exemplary embodiment of the present invention enables a domain expert (such as a client services account manager) with basic computer skills to perform all functions needed to build a new text classifier, all the way from message collection to criteria building, labeling, and deployment of a new text classifier with known performance characteristics. The tool cleverly manages message harvesting, consistent criteria development, labeling of messages, and proper machine learning protocol. It is envisioned that this end-to-end process will take less than a day instead of weeks as required by the prior art. Much of the speed-up comes in the automation of steps such as harvesting, criteria development, consistent data quality checks, and machine learning training. Some of the speed-up also comes by cleverly minimizing the number of messages that need to be labeled, which is possible because, in this exemplary embodiment, a single tool oversees both the labeling and the training of the algorithm. Some of the speed-up also comes because communications and coordination required between the different parties involved in building a prior-art classifier is removed. Only one person is necessary for building the classifier of the exemplary embodiment.

[**0008**] The present invention provides two primary advancements for this novel approach: (1) an advanced Active Learning process that combines, in the exemplary embodiment, Active Learning for training set building, relabeling for data quality and test-set building all into a single process; and (2) structured criteria elicitation, which involves a

question/answer process to a generate a clear expression of labeling criteria that is crucial in message classification.

[0009] Consequently, it is a first aspect of the current invention to provide a computer assisted/implemented method (or a computer/system or a computer memory containing software that includes instructions for implementing a method) for developing a classifier for classifying communications (text, electronic, etc.) that includes the steps of: (a) presenting communications to a user for labeling as relevant or irrelevant, where the communications are selected from groups of communications including: (i) a training set group of communications, where the training set group of communications is selected by a traditional Active Learning algorithm; (ii) a test set group of communications, where the test set group of communications is for testing the accuracy of a current state of the classifier being developed by the present method; (iii) a faulty set of communications determined to be previously mislabeled by the user; (iv) a random set of communications previously labeled by the user; and (v) a system-labeled set of communications previously labeled by the system; and (b) developing a classifier for classifying communications based upon the relevant/irrelevant labels assigned by the user during the presenting step. In a more detailed embodiment, the presenting step includes the steps of: assessing the value that labeling a set of communications from each group will provide to the classifier being developed; and selecting a next group for labeling based upon the greatest respective value that will be provided to the classifier being developed from the assessing step.

[0010] It is a second aspect of the present invention to provide a computer assisted/implemented method (or a computer/system or a computer memory containing software that includes instructions for implementing a method) for developing a classifier for classifying communications (text, electronic, etc.) that includes the steps of: (a) presenting communications to a user for labeling as relevant or irrelevant, where the communications are selected from groups of communications including: (i) a training set group of communications, where the training set group of communications is selected by traditional Active Learning algorithms; (ii) a test set group of communications, where the test set group of communications is for testing the accuracy of a current state of the classifier being developed by the present method; and (iii) a

previously-labeled set of communications previously labeled by the user, the system and/or another user; and (b) developing a classifier for classifying communications based upon the relevant/irrelevant labels assigned by the user during the presenting step. In a more detailed embodiment, the previously labeled set of communications includes communications previously labeled by the user. In a further detailed embodiment, the previously labeled set of communications includes communications determined to be possibly mislabeled by the user.

[0011] In an alternate detailed embodiment of the second aspect of the present invention, the previously-labeled set of communications may include communications previously labeled by the system. In a further detailed embodiment, the previously-labeled set of communications includes communications previously labeled by a user and communications previously labeled by the system.

[0012] It is also within the scope of the second aspect of the present invention that the presenting step includes the steps of: assessing a value that labeling a set of communications from each group will provide to the classifier being developed; and selecting the next group for labeling based upon the greatest respect of value that will be provided to the classifier being developed from the assessing step. It is also within the scope of the second aspect of the present invention that the method further includes the step of developing an expression of labeling criteria in an interactive session with the user.

[0013] A third aspect of the present invention is directed to a computer assisted/implemented method (or a computer/system or a computer memory containing software that includes instructions for implementing a method) for developing a classifier for classifying communications (text, electronic, etc.) that includes the steps of: (a) developing an expression of labeling criteria in an interactive session with the user; (b) presenting communications to the user for labeling as relevant or irrelevant; and (c) developing a classifier for classifying communications based upon the relevant/irrelevant labels assigned by the user during the presenting step. In a more detailed embodiment, the interactive session includes the steps of posing hypothetical questions to the user regarding what type of information the user would

consider relevant. In a more detailed embodiment, the hypothetical questions illicit "yes", "no" and "unsure" responses (or their equivalents) from the user. It is within the scope of the invention that the subsequent questions are based, at least in part, upon answers given to previous questions. It is also within the scope of the third aspect of the present invention that the step of developing an expression for labeling criteria produces a criteria document; where this criteria document may include a list of items that are considered relevant and a list of things that are considered irrelevant. It is also within the scope of the third aspect of the present invention that the expression and/or the criteria document include a group of key words and phrases for use by the system in automatically labeling communications. It is also within the third aspect of the present invention that the labeling step (b) includes the step of querying the user as to which items influence the label on a user-labeled communication. Finally, it is within the scope of the third aspect of the present invention that the interactive session is conducted prior to the presenting step (b).

[0014] A fourth aspect of the present invention is directed to a computer assisted/implemented method (or a computer/system or a computer memory containing software that includes instructions for implementing a method) for developing a classifier for classifying communications (text, electronic, etc.) that includes the steps of: (a) defining a domain of communications on which the classifier is going to operate; (b) collecting a set of communications from the domain; (c) eliciting labeling communication criteria from a user; (d) labeling, by the system, communications from the set of communications according, at least in part, to the labeling communication criteria elicited from the user; (e) labeling, by the user, communications from the set of communications; and (f) building a communications classifier according to a combination of labels applied to communications in labeling steps (d) and (e). In a more detailed embodiment the combination of the labeling steps (d) and (e), and the building step (f) includes the step of selecting communications for labeling by the user targeted to build the communications classifier within known performance bounds.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] Fig. 1 is a screen-shot of an initial step in an exemplary embodiment of the present invention;

[0016] Fig. 2 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0017] Fig. 3 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0018] Fig. 4 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0019] Fig. 5 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0020] Fig. 6 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0021] Fig. 7 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0022] Fig. 8 is a screen-shot of a later stage of the step of Fig. 7 according to an exemplary embodiment of the present invention;

[0023] Fig. 9 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0024] Fig. 10 is a screen-shot of a next step in an exemplary embodiment of the present invention;

[0025] Fig. 11 is a screen-shot of a next step in an exemplary embodiment of the present invention; and

[0026] Fig. 12 is a screen-shot of an example message labeled by the user according to an exemplary embodiment of the present invention.

DETAILED DESCRIPTION

[0027] The present invention is directed to a computer assisted/implemented method for developing a classifier for classifying communications (such as text messages, documents and other types of communications, electronic or otherwise). The present invention is further directed to a computer or computer system (or any similar device or collection of devices), as known or available to those of ordinary skill in the art, operating a software program including instructions for implementing such a method; or to a computer memory (resident within a computer or portable), as known or available to those of ordinary skill in the art, containing a software program including instructions for implementing such a method. While the exemplary embodiments described herein are oriented specifically toward the task of building Internet message relevancy filters, the present invention also provides a framework for building many types of communication/information classifiers.

[0028] Use of the computerized tool according to the exemplary embodiment of the present invention comprises roughly four stages, where these stages are designed to be iterative: (1) a stage defining where and how to harvest messages (i.e., from Internet message boards and the like), which also defines an expected domain of application for the classifier; (2) a guided question/answering stage for the computerized tool to elicit the user's criteria for determining whether a message is relevant or irrelevant; (3) a labeling stage where the user examines carefully-selected messages and provides feedback about whether or not it is relevant and sometimes also what elements of the criteria were used to make the decision; and (4) a performance evaluation stage where parameters of the classifier training are optimized, the best classifier is produced, and known performance bounds are calculated. In the guided

question/answering stage, the criteria are parameterized in such a way that (a) they can be operationalized into the text classifier through key words and phrases, and (b) a human-readable English criteria can be produced, which can be reviewed and edited. The labeling phase is heavily oriented toward an extended Active Learning framework. That is, the exemplary embodiment decides which example messages to show the user based upon what category of messages the system thinks would be most useful to the Active Learning process.

[0029] The exemplary embodiment of the present invention enables a domain expert (such as a client services account manager) with basic computer skills to perform all functions needed to build a new text classifier, all the way from message collection to criteria building, labeling, and deployment of a new text classifier with known performance characteristics. The tool cleverly manages message harvesting, consistent criteria development, labeling of messages, and proper machine learning protocol. It is envisioned that this end-to-end process will take less than a day instead of weeks as required by the prior art. Much of the speed-up comes in the automation of steps such as harvesting, criteria development, consistent data quality checks, and machine learning training. Some of the speed-up also comes by cleverly minimizing the number of messages that need to be labeled, which is possible because, in the exemplary embodiment, a single tool oversees both the labeling and the training of the algorithm. Some of the speed-up also comes because communications and coordination required between the different parties involved in building a prior-art classifier is removed. Only one person is necessary for building the classifier of the exemplary embodiment.

[0030] The present invention provides two primary advancements for this novel approach: (1) an advanced Active Learning process that combines, in the exemplary embodiment, Active Learning for training set building, relabeling for data quality and test-set building all into a single process; and (2) structured criteria elicitation, which involves a question/answer process to a generate a clear expression of labeling criteria that is crucial in message classification.

[0031] Advanced Active Learning

[0032] The advanced Active Learning process combines, in the exemplary embodiment, Active Learning for training set building, relabeling for data quality, and test set building all into a single process. During the labeling process, the tool chooses which messages, sets of messages and/or categories of messages to present to the human labeler by balancing the relative importance of the above three types of labeling (training set building, relabeling and test set building). More specifically, the exemplary embodiment of the tool chooses between five different labeling categories of messages that may be selectively presented to the human labeler based upon the greatest respective value that labeling messages of the respective category will provide to the classifier being developed during this process. These five different types of labeling categories are as follows: (1) a training set group of messages, where the training set group of messages is selected by a traditional Active Learning algorithm; (5) a system-labeled set of messages previously labeled by the tool used to augment the training set while training the text classifier; (3) a test set group of messages, where the test set group of messages is used for testing the accuracy of a current state of classifier being developed; (4) a faulty set of messages suspected by the system to be previously mislabeled by the user; and (5) a random set of messages previously labeled by the user used to estimate how error-prone the human labeler may be.

[0033] The Training Set Group of Messages. The traditional Active Learning algorithm selects messages/examples that, along with their user-provided label, will help the classifier do a better job classifying in the future. There are many selection criteria available in the literature, and they include things like: picking a message about which the classifier is very uncertain, picking a message that is similar to many other messages, picking a message that statistically is expected to teach a lot, etc.

[0034] The System-Labeled Set of Messages. The system-labeled set of messages, which have been previously automatically labeled by the classifier, may be provided to the human labeler to see if the tool needs to correct any errors in the automatic key word matching

labeling process. The key words are automatically derived from the criteria elicitation process discussed below. The tool currently seeds the training phase of the exemplary embodiment with a set of example messages that have been automatically labeled by simple key word matching. This often provides a good starting point, but there are going to be mistakes in the key word labeling. By presenting these to the human labeler for review, the tool can correct any errors here.

[0035] The Test Set Group of Messages. The test set group of messages is a randomly-chosen test set example. This set will be used to evaluate how the current classifier is performing. More precisely, statistical confidence bounds can be placed on the current accuracy, Precision/Recall Break Even, F-1 or other performance measures of the classifier. It is desired to maximize the 95% confidence lower bound of the classifier. By adding more test set examples, the system allows the region of confidence to be tighter, which raises the lower bound on performance. For example, if a classifier is performing at 80% ± 5%, processing a new test set message may be found to improve the variance to 80% ± 3%.

[0036] The Faulty Set of Messages. This set of messages is essentially a bad-looking example previously shown to the user. This set is based upon the understanding that there are almost always inconsistencies with human labeling of communications. These inconsistencies can be very damaging to some classification algorithms. Some of these inconsistencies are easy to spot by the tool. For example, a communication that the classifier thinks is relevant but the human labeler labeled as irrelevant may often-times be a labeling mistake. By showing these examples again to the user, the tool can correct some of these mistakes and improve the classification.

[0037] The Randomly-Selected Set of Messages. The randomly-selected set of messages, which have been previously labeled by the human labeler, may be provided to the human labeler for labeling again to estimate how consistent the labeler is labeling messages. By understanding how consistent the labeling is being conducted by the labeler, the tool will know how aggressively to try to correct labeling. In turn, by showing some randomly-selected

examples, the tool can judge how frequently it should show sets of communications that it determines are likely to be faultily labeled communications for relabeling.

[0038] Recognizing that labeling the above-discussed five categories of messages is valuable, the next determination for the system is when to send a particular category of messages to the human labeler and in what proportions. This is determined by mathematically expressing (in terms of improvement to expected lower bound on measured performance of the classifier) the additional value for labeling each category of messages. This will give the tool a priority for presenting each category of messages to the user for labeling. Of course, these priorities will change over time. For example, when just starting out, it is more important to label test sets of messages, because without labeling test sets the system cannot measure the overall performance. After some time, the test set will be large enough that adding to it is less important, and at this point, it is likely that other categories of labels will become relatively more important. In its simplest form, the rates of labeling from the different sets of messages can just be fixed to set percentages. This does not give optimal performance, but it is computationally easier.

[0039] Labeling an additional Test Set message increases the expected lower bound on measured performance by making the error bars on the expectation smaller because the error would be measured over a larger set of data. The value of labeling such a message can be calculated by the expected decrease in the size of the error bars.

[0040] Labeling an additional Training Set message increases the expected lower bound on measured performance by improving the expected measured performance because it provides an additional training example to the learning algorithm. The value of labeling such a message could be calculated by measuring the expected gain in performance as predicted by the active labeling algorithm. It could also be calculated by measuring the slope of the learning curve as more data is labeled.

[0041] Labeling a Faulty message increases the expected lower bound on measured performance by improving the expected measured performance because it changes the label of a

training (or test) example that was proving difficult for the classifier to incorporate. The value of labeling such a message can be calculated by measuring the improvement in classifier performance if the label were changed, multiplied by the probability the label will be changed, as estimated from the number of labeling changes from previously labeling Faulty messages and Randomly Selected messages.

[0042] Labeling a System-Labeled message increases the expected lower bound on measured performance by improving the expected measured performance because sometimes it will correct the label assigned by the system. The value of labeling such a message could be calculated by measuring the improvement in classifier performance if the label were changed, multiplied by the probability the label will be changed, as estimated from the frequency that previously-labeled System-Labeled messages have had their label changed.

[0043] Labeling a Randomly-Selected message indirectly increases the expected lower bound on measured performance. The value of labeling such a message lies in accurately estimating the error rate, which determines how aggressively to label Faulty messages. The rate of which Randomly-Selected messages are labeled can be calculated using the lower-bound on the expected frequency that Faulty messages get their labeling changes.

[0044] Consequently, it is a first aspect of the current invention to provide a computer assisted/implemented method (or a computer/system or a computer memory containing software that includes instructions for implementing a method) for developing a classifier for classifying communications that includes the steps of: (a) presenting communications to a user for labeling as relevant or irrelevant, where the communications are selected from groups of communications including: (i) a training set group of communications, where the training set group of communications is selected by a traditional Active Learning algorithm; (ii) a system-labeled set of communications previously labeled by the system; (iii) a test set group of communications, where the test set group of communications is for testing the accuracy of a current state of the classifier being developed by the present method; (iv) a faulty set of communications suspected by the system to be previously mislabeled by the user; and (v) a random set of communications

previously labeled by the user; and (b) developing a classifier for classifying communications based upon the relevant/irrelevant labels assigned by the user during the presenting step. In a more detailed embodiment, the presenting step includes the steps of: assessing the value that labeling a set of communications from each group will provide to the classifier being developed; and selecting a next group for labeling based upon the greatest respective value that will be provided to the classifier being developed from the assessing step.

[0045] It is a second aspect of the present invention to provide a computer assisted/implemented method (or a computer/system or a computer memory containing software that includes instructions for implementing a method) for developing a classifier for classifying communications that includes the steps of: (a) presenting communications to a user for labeling as relevant or irrelevant, where the communications are selected from groups of communications including: (i) a training set group of communications, where the training set group of communications is selected by traditional Active Learning algorithms; (ii) a test set group of communications, where the test set group of communications is for testing the accuracy of a current state of the classifier being developed by the present method; and (iii) a previously-labeled set of communications previously labeled by the user, the system and/or another user; and (b) developing a classifier for classifying communications based upon the relevant/irrelevant labels assigned by the user during the presenting step. In a more detailed embodiment, the previously labeled set of communications includes communications previously labeled by the user. In a further detailed embodiment, the previously labeled set of communications includes communications determined to be possibly mislabeled by the user.

[0046] In an alternate detailed embodiment of the second aspect of the present invention, the previously-labeled set of communications may include communications previously labeled by the system. In a further detailed embodiment, the previously-labeled set of communications includes communications previously labeled by a user and communications previously labeled by the system.

[0047] It is also within the scope of the second aspect of the present invention that the presenting step includes the steps of: assessing a value that labeling a set of communications from each group will provide to the classifier being developed; and selecting the next group for labeling based upon the greatest respect of value that will be provided to the classifier being developed from the assessing step. It is also within the scope of the second aspect of the present invention that the method further includes the step of developing an expression of labeling criteria in an interactive session with the user. This will be described in further detail below.

[0048] Structured Criteria Elicitation

[0049] Structured criteria elicitation is based upon the idea that a clear expression of labeling criteria is crucial in a message classification process. By enforcing an elicitation stage before the labeling stage, the exemplary embodiment can make sure that the user has clearly defined in their mind (and to the tool) what they mean by relevant and irrelevant documents/messages/communications. The exemplary embodiment of the present invention provides a novel and interesting way to conduct this efficiently, and it is a powerful technique for ensuring that the labeling process proceeds smoothly and gives consistent results.

[0050] The exemplary embodiment defines a structured formalism in the message relevancy domain that guides the criteria elicitation. A full relevancy criteria is viewed as a series of bullet items. Each bullet item is a tuple: [product; aspect; strength; relevancy; key words]. To give a simple example:

> The tuple representing the concept "any message discussing the Nissan 350Z Charity Auction is relevant" is: [Nissan 350Z; corporate activity; discussions and opinions; irrelevant, "charity auction"]

[0051] By viewing labeling criteria bullet items as a point in a structured domain, specifying a labeling criteria then becomes a search for the separator (between relevant and

irrelevant communications) in the space of all criteria. By cleverly posing hypothetical questions to the user during criteria elicitation, the exemplary embodiment of the present invention can efficiently search this space and construct the criteria specification automatically from a set of "yes/no/unsure" questions posed to the user. During this process the user also supplies key words and phrases with each criteria specific dimension. As introduced above, in addition to adding to the criteria specification, such keywords may also be utilized by the system to collect groups of Internet messages using a keyword Web search during an initial message collection stage.

[0052] For internet messages about a specific consumer product, we have discovered that most labeling criteria can be expressed with several structured dimensions. The first dimension is which product is being discussed. This could be the product (such as the Nissan 350z) or a set of competitors (such as the Honda S2000). The second dimension is the aspect being discussed for the selected product. This could be a feature of the product (such as the headlights), corporate activity by the product's company, advertising about the product, etc. The third dimension is what type of discussion or mention of the product and aspect is occurring. The weakest discussion is a casual mention of the product. A stronger mention is a factual description of the product. An even stronger mention is a stated opinion of the product or a comparison of the product to its competitors. Relevance criteria specify a certain strength of discussion for each aspect of a product that is required to make it relevant.

[0053] We believe that most relevance criteria, even those for other text classification tasks, can be specified in this multi-dimensional way with the appropriate set of dimensions. By posing these criteria in this multi-dimensional way, a structured questionnaire will efficiently elicit the criteria from the human.

[0054] In the exemplary implementation of the invention, Internet message relevancy filters for marketing analysis, the first dimension (the topic) question segment is either:

- "the product"

- "the competitors"

In the exemplary embodiment, we often ignore the differentiation between the product and the competitors. The second dimension (the aspect of the topic) question segment is either:

- "a feature of the product"
- "the product itself"
- "corporate activity by the company"
- "the product's price"
- "a news article mentioning the product"
- "advertising for the product"

The third dimension (the type of discussion) question segment is either:

- "a casual mention of"
- "a factual description of"
- "a usage statement about"
- "a brand comparison involving"
- "an opinion about"

[0055] The questionnaire, in the exemplary embodiment, is built using combinations of terms taken from the three dimensions introduced above. For example, the question: "Is a brand comparison involving corporate activity by the company of the competitors relevant, irrelevant or are you unsure?" is built using the third dimension (type of discussion) segment "a brand comparison involving", the second dimension (aspect of the topic) segment "corporate activity by the company" and the first dimension (topic) segment "the competitors". Some combinations do not make sense for every aspect. For example, it does not really make sense to build a question about: "a usage statement about corporate activity by the company". Consequently, in the exemplary embodiment, the following second and third dimension combinations are permitted:

Second Dimension	Permitted Third Dimension
a feature	MENTION, DESCRIPTION, USAGE, COMPARISON, OPINION
the product itself	MENTION, DESCRIPTION, USAGE, COMPARISON, OPINION
corporate activity	DESCRIPTION, COMPARISON, OPINION
Price	DESCRIPTION, COMPARISON, OPNION
News	MENTION, DESCRIPTION, OPINION
Advertising	MENTION, DESCRIPTION, COMPARISON, OPINION

[0056] In the exemplary embodiment, criteria elicitation is a questionnaire, where the later questions are created based upon the answers to the earlier questions. For example, one early question might be, "Is a factual description of a feature of the product relevant?". If the answer is no, a follow-up question might be, "Is an opinion about a feature of the product relevant?". If the answer is yes, a more appropriate question would be, "Is a casual mention of a feature of a product relevant?". Basically, each question builds upon the previous one, pushing the boundaries until the system sees a cross-over from relevancy or irrelevancy or vice-versa.

[0057] The end result of the user answering the questions provided by the questionnaire is a criteria document, which is a human-readable bulleted list defining the types of things that are relevant and the types of things that are irrelevant. This document is good for external review. The document is also used inside the tool. The key words defined for each bullet item help pre-seed what types of phrases to look for in the feature extraction. They are also used to pre-label some examples based on key word and phrase matching. During labeling, the tool may periodically ask the user to identify which bullet items were used to label a specific example. This can be used to refine the set of key words, and also to ensure the consistency of the labeling by the user.

[0058] Additionally, with the exemplary embodiment, after the questionnaire is provided to the user, the user is given the opportunity to add new values for the second dimension, although it has been found that this does not occur very often.

[0059] Consequently, it can be seen that a third aspect of the present invention is directed to a computer assisted/implemented method for developing a classifier for classifying communications that includes the steps of: (a) developing an expression of labeling criteria in an interactive session with the user; (b) presenting communications to the user for labeling as relevant or irrelevant; and (c) developing a classifier for classifying communications based upon the relevant/irrelevant labels assigned by the user during the presenting step. In a more detailed embodiment, the interactive session includes the steps of posing hypothetical questions to the user regarding what type of information the user would consider relevant. In a more detailed embodiment, the hypothetical questions elicit "yes", "no" and "unsure" responses (or their equivalents) from the user. It is within the scope of the invention that the subsequent questions are based, at least in part, upon answers given to previous questions. It is also within the scope of the third aspect of the present invention that the step of developing an expression for labeling criteria produces a criteria document; where this criteria document may include a list of items that are considered relevant and a list of things that are considered irrelevant. It is also within the scope of the third aspect of the present invention that the expression and/or the criteria document include a group of key words and phrases for use by the system in automatically labeling communications. It is also within the third aspect of the present invention that the labeling step (b) includes the step of querying the user as to which items influence the label on a user-labeled communication. Finally, it is within the scope of the third aspect of the present invention that the interactive session is conducted prior to the presenting step (b).

[0060] EXAMPLE END-TO-END PROCESS

[0061] The following is an example of a graphical process provided by an exemplary embodiment of the present invention to build a new text classifier using the advanced active learning and the structured criteria elicitation processes discussed above.

[0062] As shown in Fig. 1, a first step is to query the user of the project name. This project name will be used to later identify the structured criteria document and other related materials.

[0063] As shown in Fig. 2, a next step is to request the user to specify a variety of data feeds or sources from which the system will harvest the data. These sources will be used during both training and production. The data sources may be a collection of Internet message or news group messages (or other alternate communications, such as emails, chat room discussions, instant messenger type discussions and the like) previously collected and stored at the specified location, and/or may be the locations (such as Web or NNTP addresses or links) from which messages will be harvested.

[0064] As shown in Fig. 3, a next step is to have the user enter a set of phrases that identify, describe or are associated with the general type of product being searched. This is used to define a product category that the present project will focus on.

[0065] As shown in Fig. 4, a next step is to request the user to enter a set of phrases that name the customer and their product. These phrases can include specific brand names, for example.

[0066] As shown in Fig. 5, a next step is to request the user to enter a set of phrases that name competing companies and branded products relevant to the present project.

[0067] As shown in Fig. 6, a next step would be to request the user to enter counter-example phrases that indicate a particular communication is not related to the key concept. For example, in the present example, the user may enter the brand names of popular video game consoles and associated street racing games to eliminate messages that discuss the relevant automobile product in reference to its use in a video game.

[0068] Fig. 7 provides an example of a criteria questionnaire, which asks specific questions to the user whether certain criteria would be "relevant/irrelevant/unsure". For example, as shown in Fig. 8, a brand comparison involving the product itself is considered relevant, a brand comparison involving a feature of the product is considered relevant but a factual description of corporate activity by the company is irrelevant. As discussed above, the specific answers to each of these criteria questions is used by the exemplary embodiment to develop subsequent questions that build upon the answers to the previous questions.

[0069] As shown in Fig. 9, the answers provided by the user to this questionnaire will be used to build a set of labeling criteria. This set of labeling criteria is used so that the user can verify the labeling criteria that was defined as a result of the questionnaire and to also refine the labeling criteria. As introduced above, at this stage, the user is given the opportunity to add keywords to each criteria element to enhance the tool's performance. This refinement can involve adding key words to each criteria element, changing the relevancy or tone of the criteria statements or deleting any statement entirely.

[0070] As shown in Fig. 10, the present exemplary embodiment will save the human-readable criteria statements into a criteria document. As discussed above, this criteria document can help the user verify to himself or herself at any time what he or she originally considered relevant so that subsequent labeling operations can be consistent; and further, the criteria statements are also utilized by the system in automatic labeling.

[0071] As shown in Fig. 11, a next step in the exemplary embodiment is to allow the user to begin labeling messages according to the advanced active learning process introduced above. Specifically, the tool chooses which messages, sets of messages and/or categories of messages to present to the human labeler by balancing the relative importance of the above three types of labeling (training set building, relabeling and test set building). Fig. 12 provides an example of a message to be labeled by the user. As can be seen in Fig. 12, certain key words have been highlighted by the system to give a user a more specific idea of why the system considered this message to be in need of labeling.

[0072] Following from the above description and invention summaries, it should be apparent to those of ordinary skill in the art that, while the systems and processes herein described constitute exemplary embodiments of the present invention, it is understood that the invention is not limited to these precise systems and processes and that changes may be made therein without departing from the scope of the invention as defined by the following claims. Additionally, it is to be understood that the invention is defined by the claims and it is not intended that any limitations or elements describing the exemplary embodiments set forth herein are to be incorporated into the meanings of the claims unless such limitations or elements are explicitly listed in the claims. Likewise, it is to be understood that it is not necessary to meet any or all of the identified advantages or objects of the invention disclosed herein in order to fall within the scope of any claims, since the invention is defined by the claims and since inherent and/or unforeseen advantages of the present invention may exist even though they may not have been explicitly discussed herein.

[0073] What is claimed is: